Gramener Case Study

Group Members

- Kumar Abhilash
- Barsha Guha
- Jaya Rathina
- Rakhee Roy

Strategy and Business Objectives

Business Objective

- Identify how consumer attributes and loan attributes influence the tendency of default.
- **Strategy**
- Identify patterns which indicate if a person is likely to default which will help Gramener in reducing credit losses for company.
- Understand the driving factors (or driver variables) behind loan default.
- Gramener can utilize this knowledge for its portfolio and risk assessment for future loan requests.

Data - Loan Data (2007 – 2011)

Data

- Loan data contain all loans issued through the time period 2007 to 2011
- 39718 application during the mentioned tenure with 111 variables

Analysis Steps

The analysis is divided into below steps:

- Business Understanding -> Data Understanding
- Data exploratory analysis involving data cleaning.
- Identify driving variables by performing below steps

- Uni-variate Analysis for categorical variable
- Segmented Uni-variate analysis
- Uni-variate Analysis for numeric variables
- Bi-variate Analysis for continuous variable

Analysis Steps ..contd..

- Bi-variate Analysis for categorical variable.

- Identify all important trend that plots are conveying and driving variable.

- Make a PPT to present your analysis to the chief data scientist of your company.

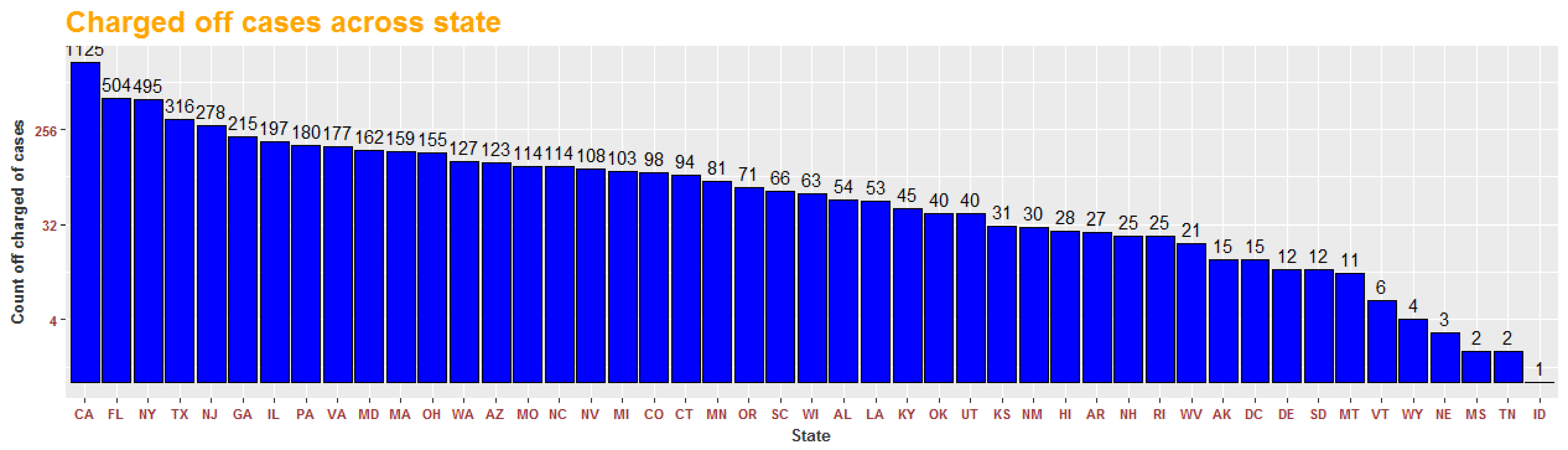
- In the chosen funding type, which **countries** have witnessed the most funding?

- In the chosen funding type and top countries, which **sectors** have performed the best?

Data cleaning Steps

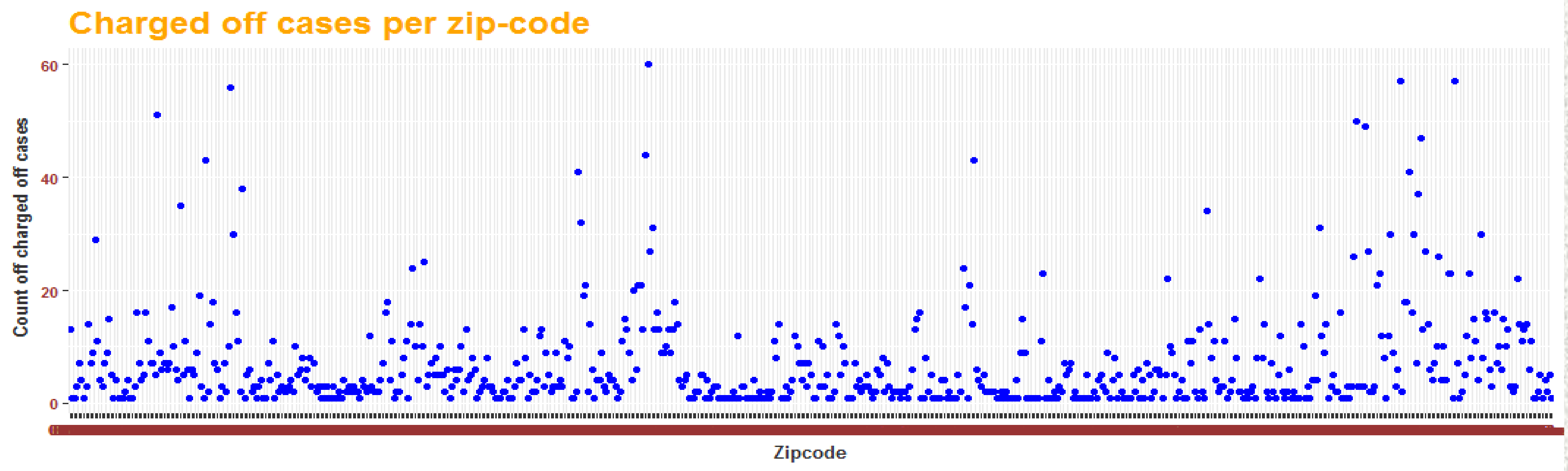
✓ Remove all variables that only contain one value or two values.
✓Remove all variables which has more than 60 percent NA values.
✓Identify all columns that don’t provide any meaningful data for analysis like URL.
✓Since we are not going to do text analysis, those columns are useless like “desc”.
✓Remove redundant columns like “title” since purpose also captures same information.
✓Loan id and member id are unique and have one to one mapping , so drop one column among them i.e member_id
✓ Some columns like “int_rate” , “revol_util” are characters due to presence of %. Convert them to numeric.
✓Convert variable term to numeric value by removing “ months ” from the end.
✓Verify the dates are imported correctly. Covert them to correct format.

Univariate Analysis Plot-1



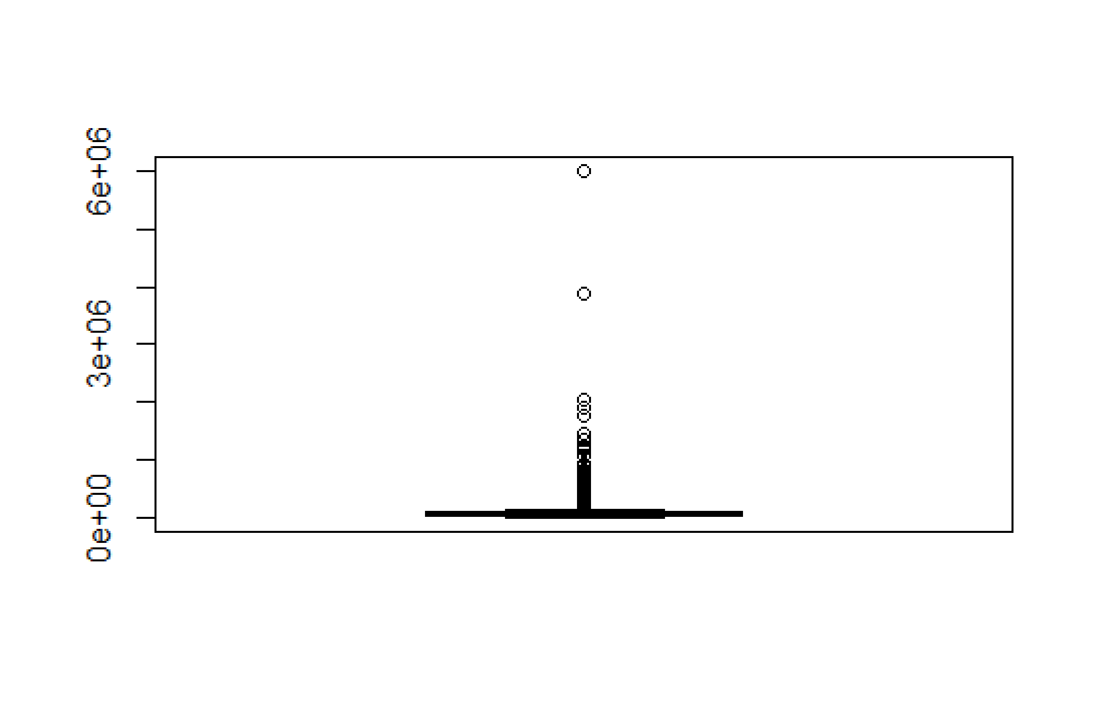
Top states with charged off cases are California ,Florida ,Newyork ,Texas, New Jersey.

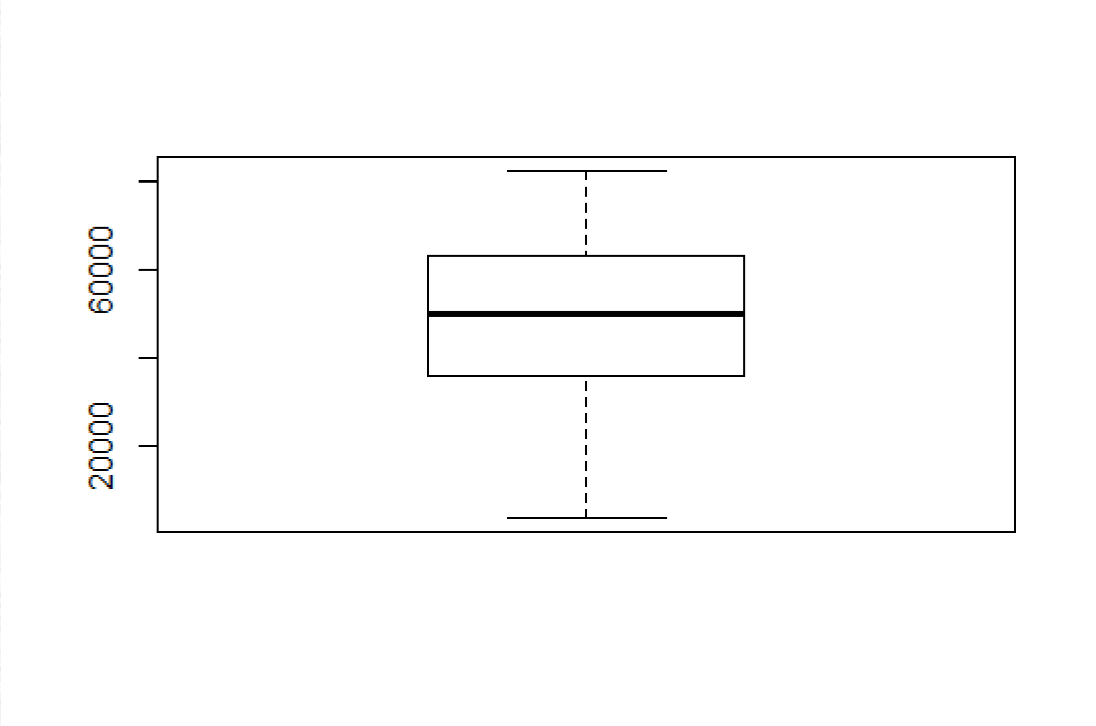
Uni-variate Analysis Plot-2



While plotting we arranged zip codes in ascending order. So from the plot it can interpreted that most of the charged off cases are in 9XXXX which is the state of California.

Uni-variate Analysis Plot-3 –Annual Income Distribution distribution



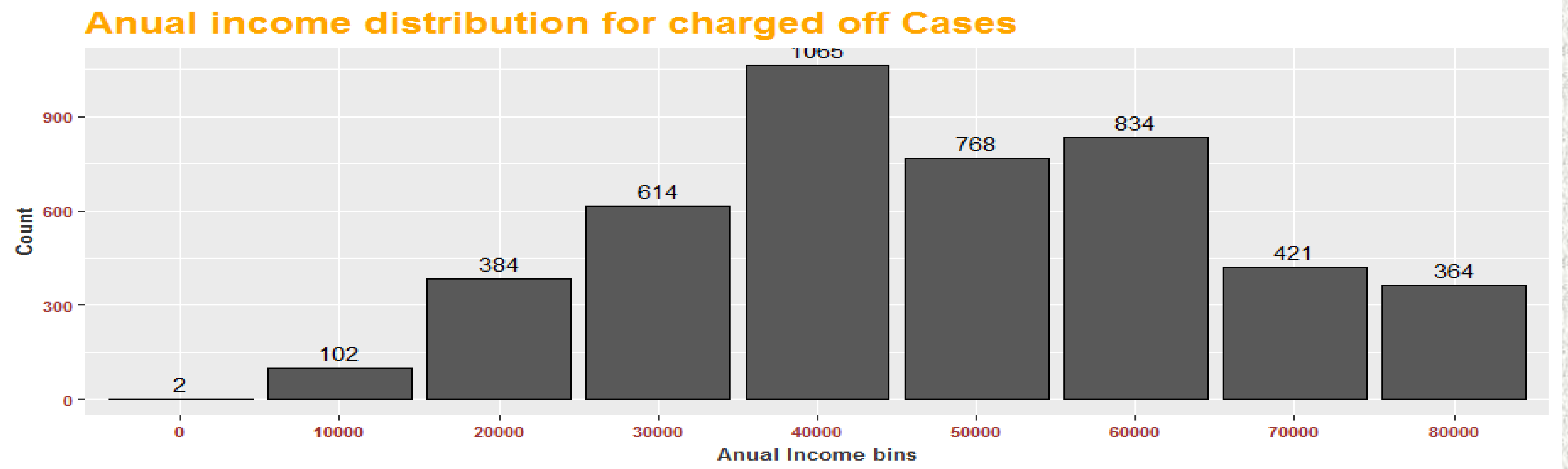


```
summary(loan$annual_inc)
Min. 1st Qu. Median Mean 3rd Qu. Max.
4000 40000 58868 68778 82000 6000000
```

```
summary(loan$annual_inc)
 Min. 1st Qu. Median Mean 3rd Qu. Max.
 4000 36000 50000 49532 62892 82000
```

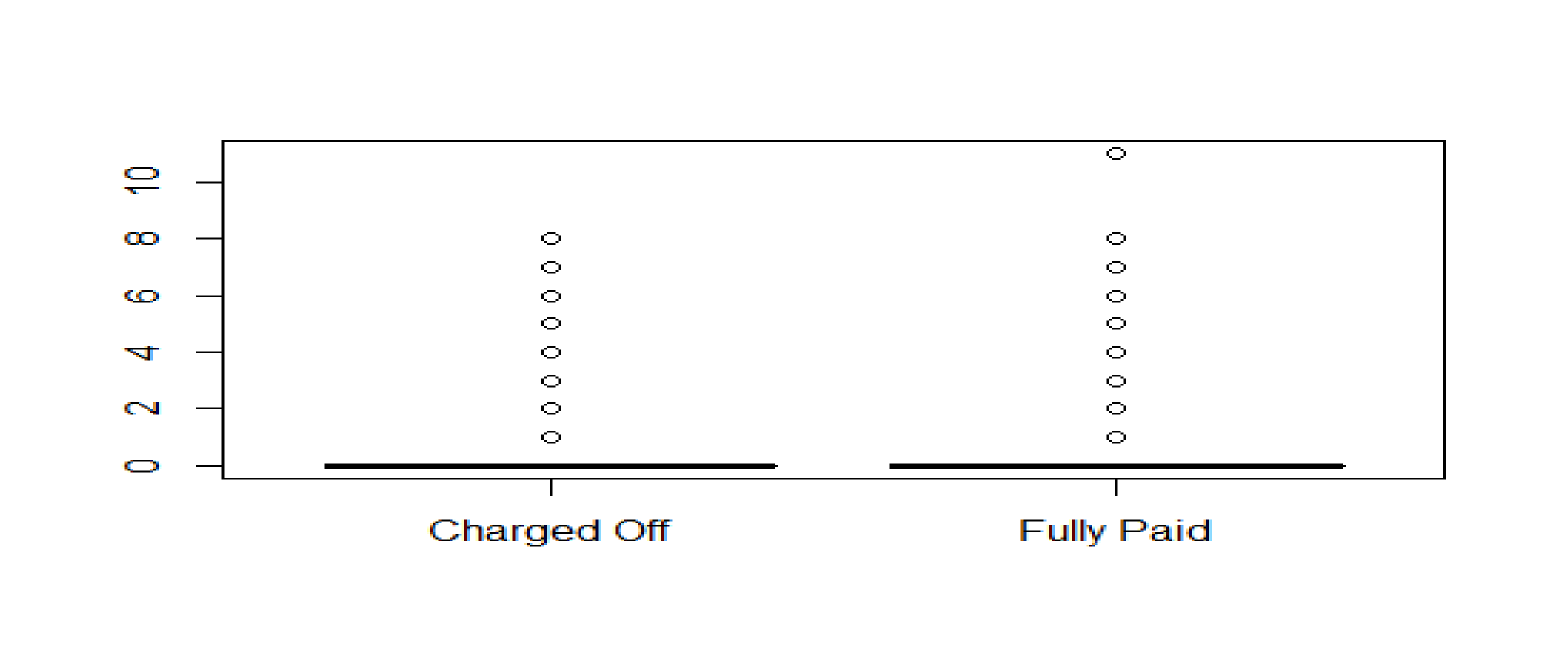
We can see in first plot that there are many outliers. So we removed outliers by taking values less than 3rd Qu. which is 82000.After cleaning , distribution of annual income looks good.

Contd..Plot-3 –Annual Income Distribution



Dividing annual income into bins of 10000’s, we can see that income bin having values between 40000 and 50000 have highest charged of cases.

Plot-4 Delinq_2yrs analysis

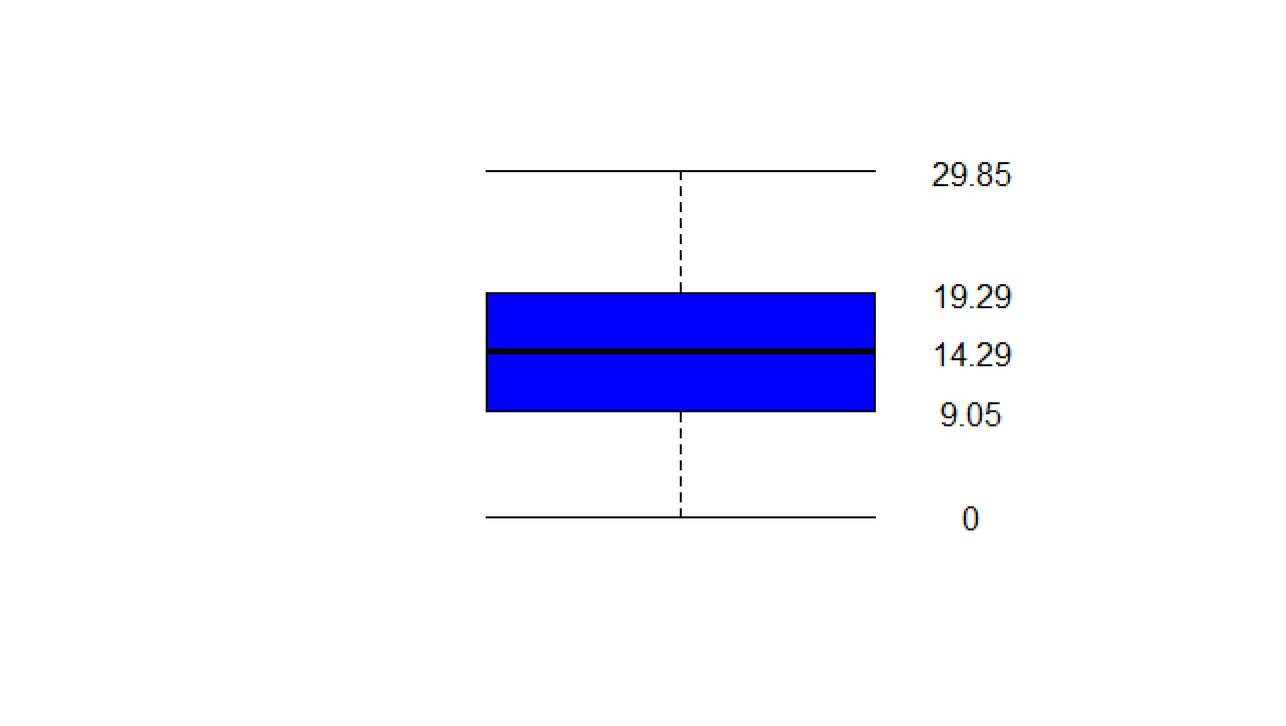


```
summary(loan$delinq_2yrs)
Min. 1st Qu. Median Mean 3rd Qu. Max.
0.0000 0.0000 0.0000 0.1383 0.0000 11.0000

nrow(loan_Charged_off[loan_Charged_off$delinq_2yrs==0,])
[1] 4936
nrow(loan_FP[loan_Charged_off$delinq_2yrs==0,])
[1] 28899
 nrow(loan_Charged_off)
 [1] 5627
nrow(loan_FP)
 [1] 32950
```

Number of zero values are more in both loan statuses i.e Fully paid, charged off

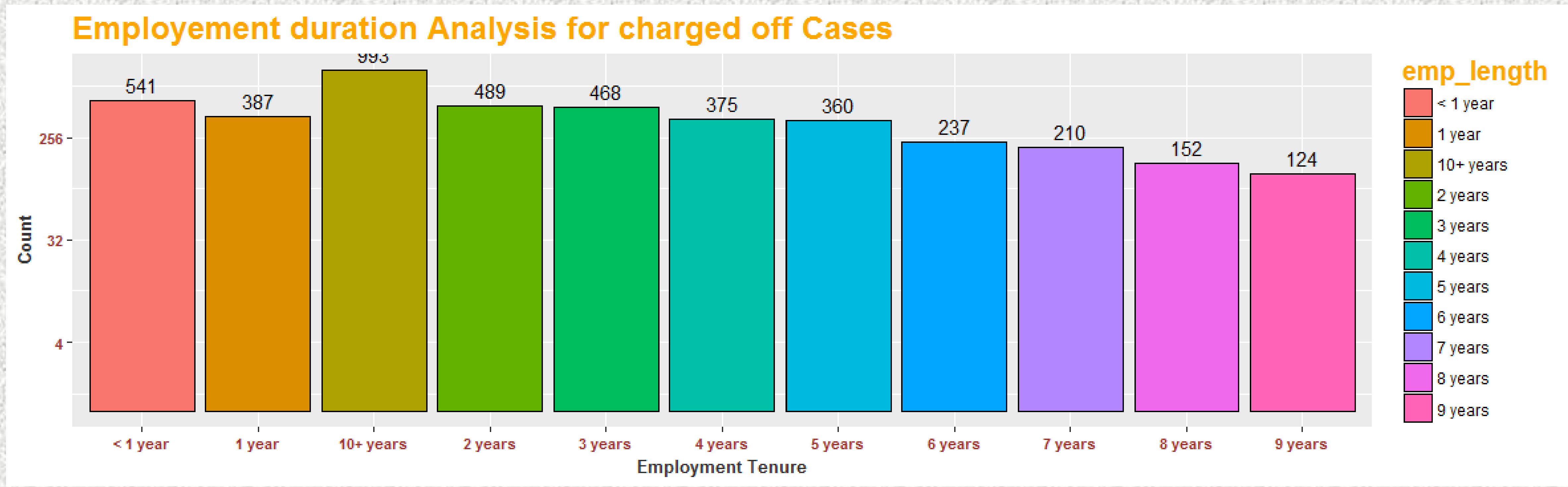
Plot-5 DTI analysis for charged of cases



```
summary(loan_Charged_off$dti)
Min. 1st Qu. Median Mean 3rd Qu. Max.
 0.00 9.05 14.29 14.00 19.29 29.85
```

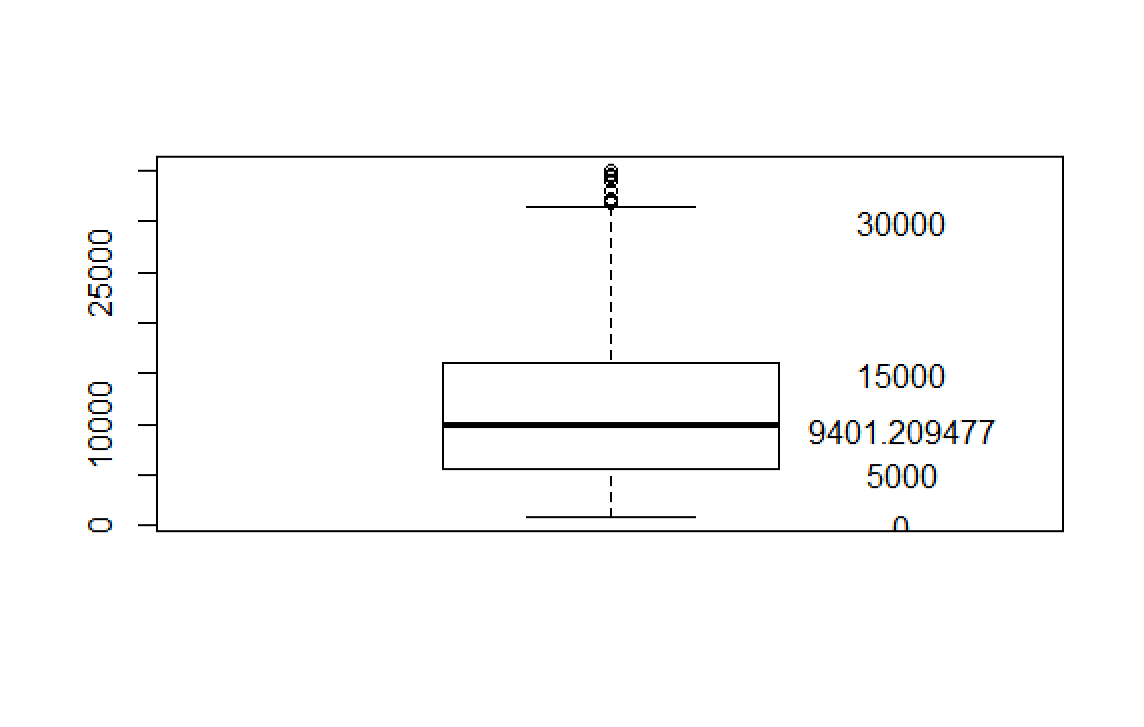
DTI Plot seems to be well distributed

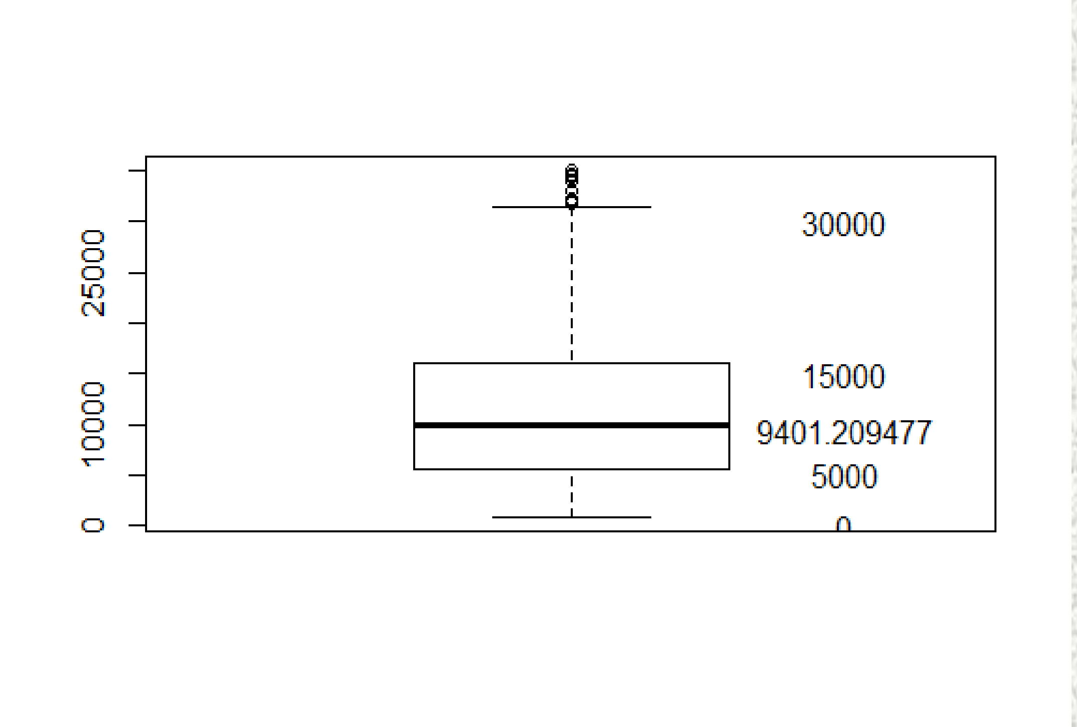
Plot-6 Employment duration Analysis



Decreasing trend with employment tenure can be seen from plot. Maximum number of charged of cases are seen in tenure less than 1 year. Tenure with more than 10 year has also high number of charged of cases because it is accumulation of all tenure length > 10.

Plot-7 Funded amount and Fund amount invested



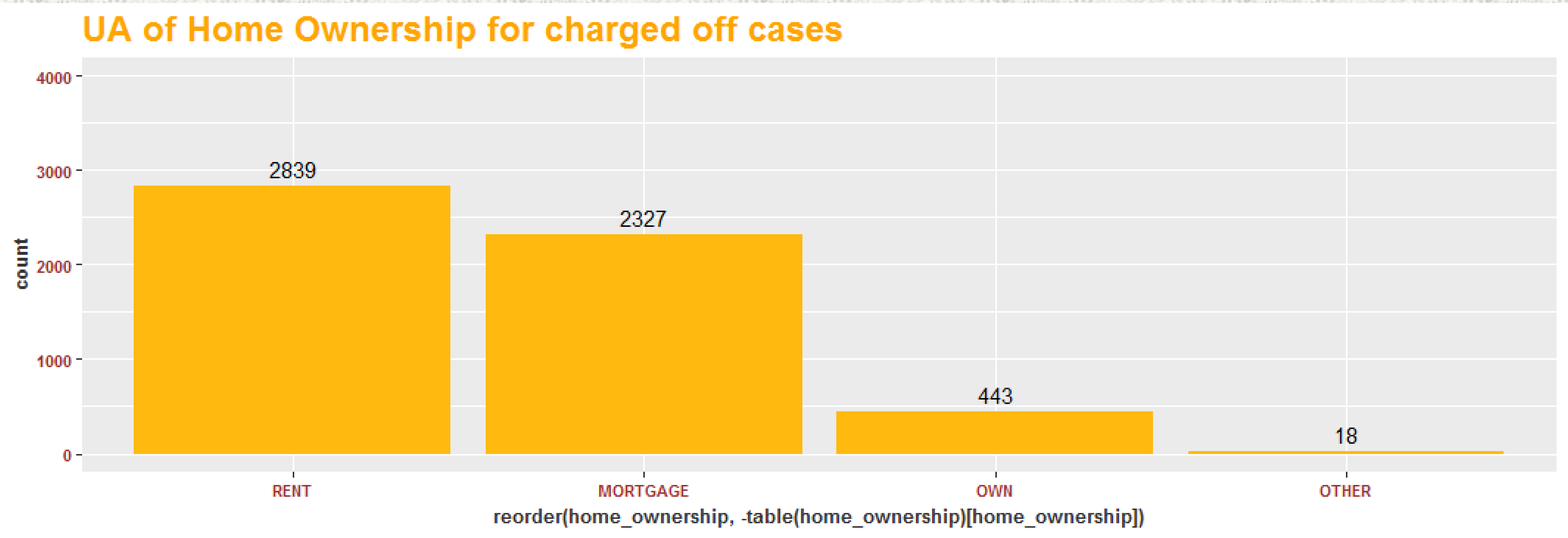


```
summary(loan_Charged_off$funded_amnt)
Min. 1st Qu. Median Mean 3rd Qu. Max.
 900 5575 10000 11753 16000 35000
```

```
summary(loan_Charged_off$funded_amnt_inv)
Min. 1st Qu. Median Mean 3rd Qu. Max.
 0 5000 9401 10865 15000 35000
```

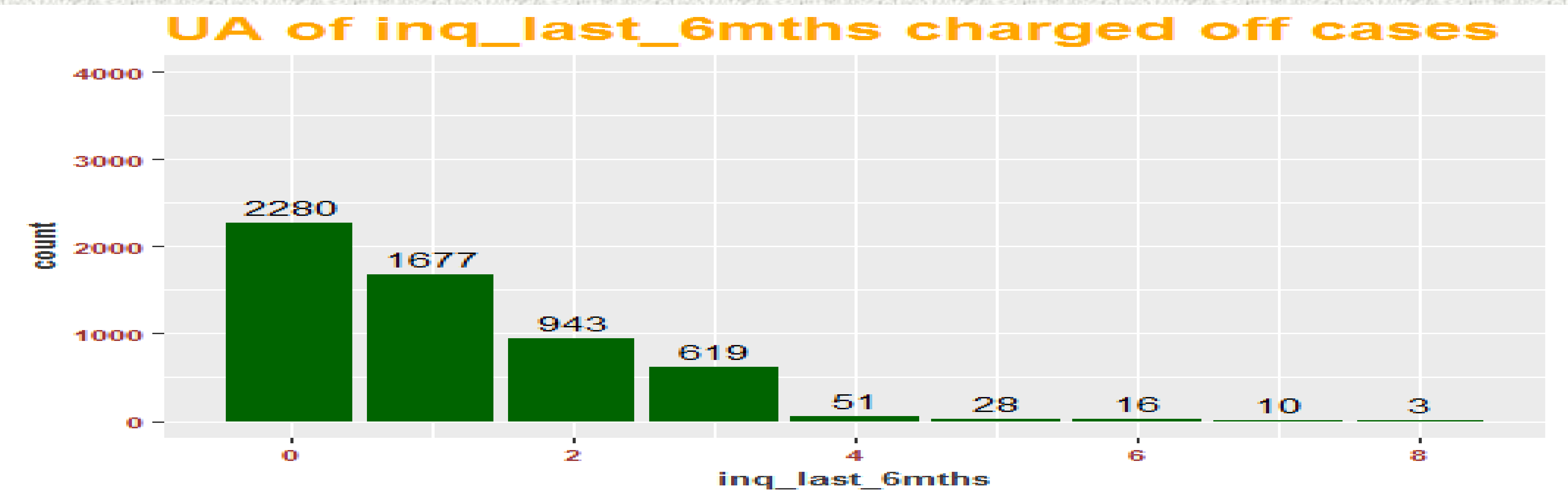
This is the plot for charged of cases. Many outliers are seen which needs to be removed for better analysis.

UA-Plot-8 Home Ownership for charged off



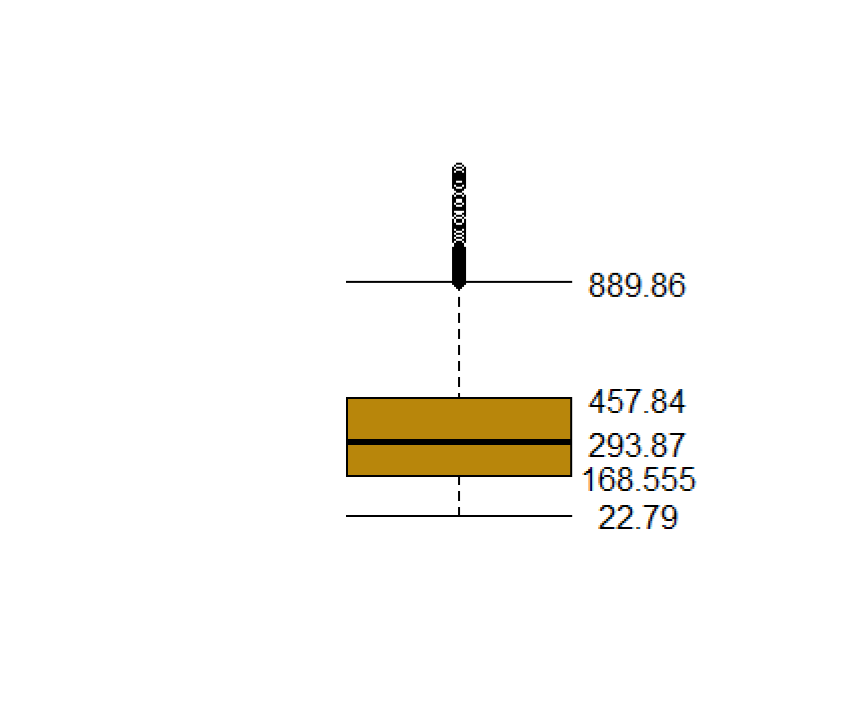
Charged off cases are maximum in case of rented houses applicants followed by Mortgage.

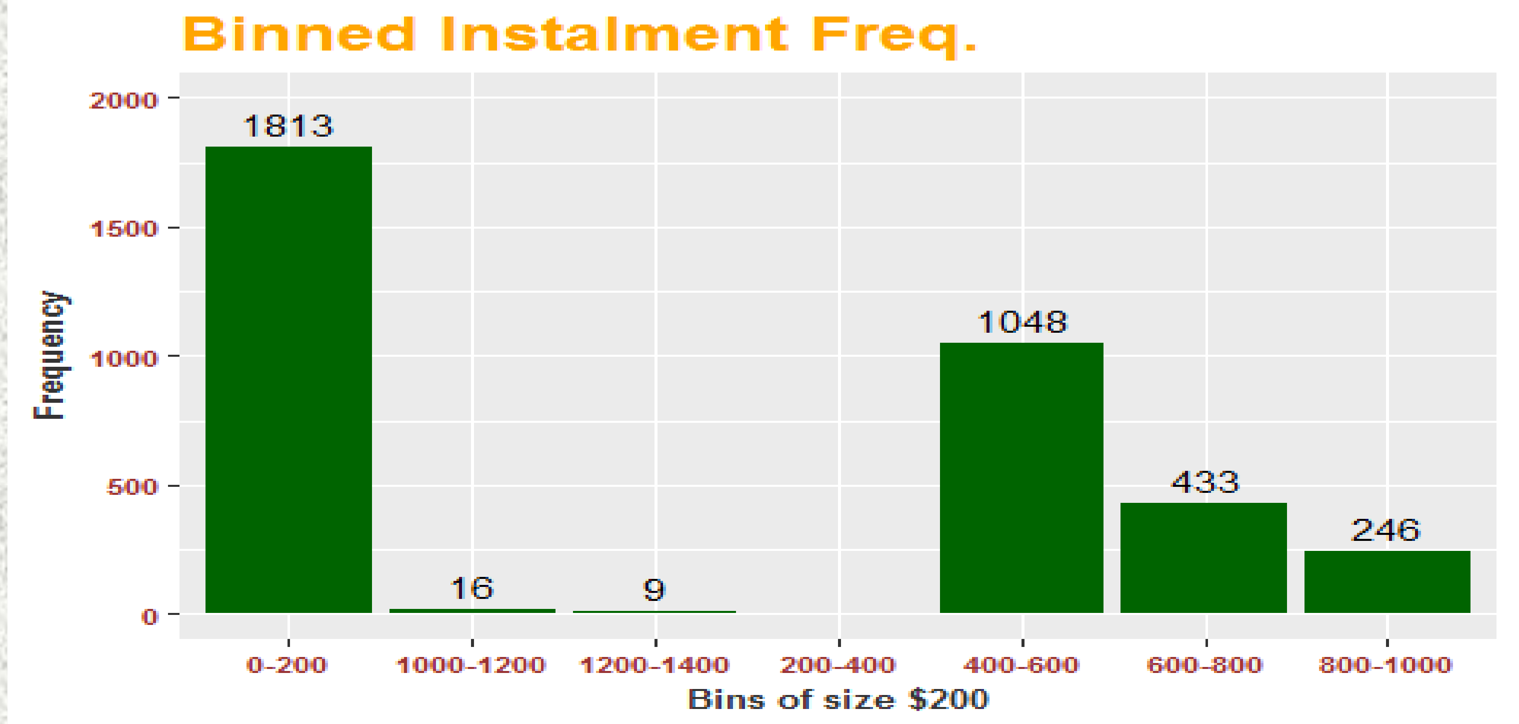
UA-Plot-9 Inquiry in last 6 months for charged off cases



From the plot it is evident that maximum rise is for zero value which clearly shows that
Maximum no of charged off cases doesn't have any inquiries in last 6 months.

UA-Plot-10 Installment analysis in charged off cases

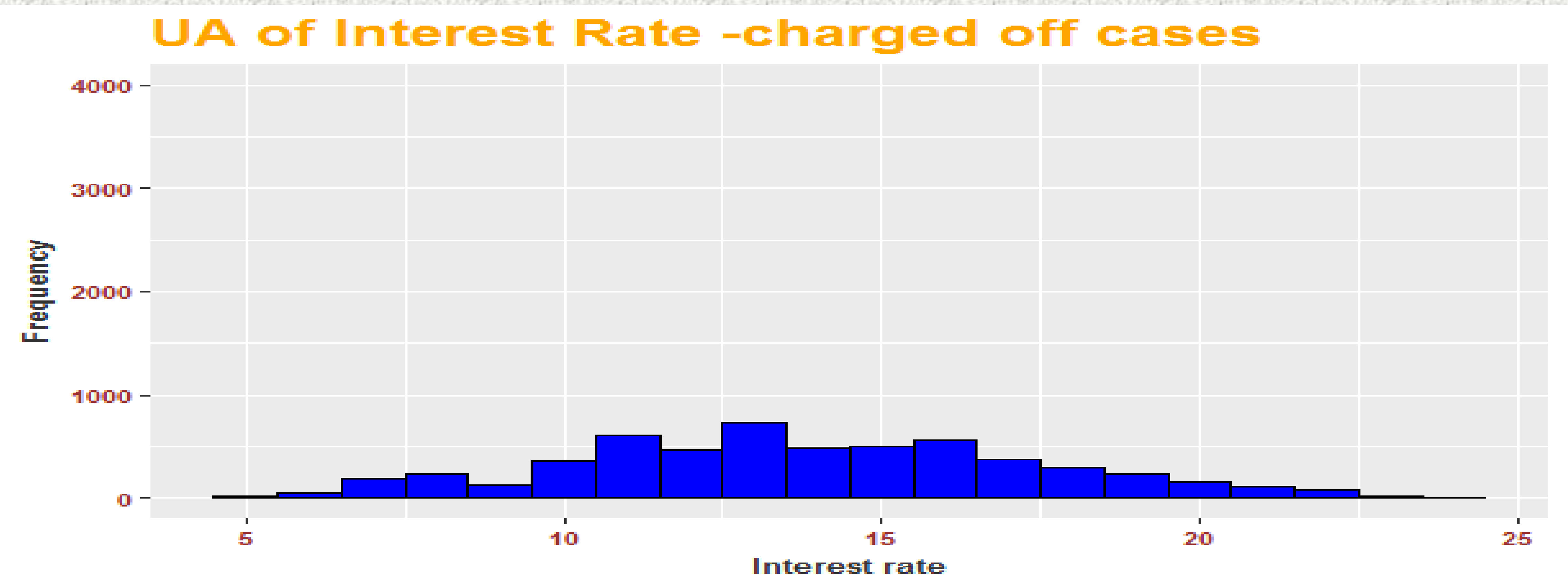




```
summary(loan_Charged_off$installment)
Min. 1st Qu. Median Mean 3rd Qu. Max.
22.79 168.56 293.87 336.18 457.84 1305.19
```

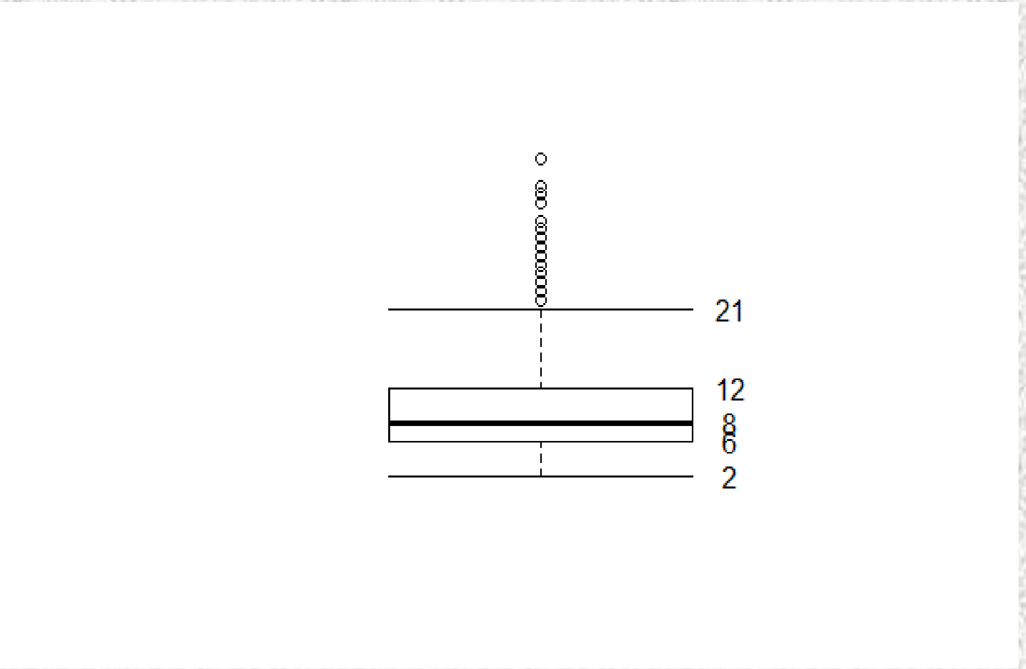
From box plot we can see that there are many outliers which needs to be removed. After dividing installment into bin of 200's, we see that bins with higher installment ranges have less charged off cases.

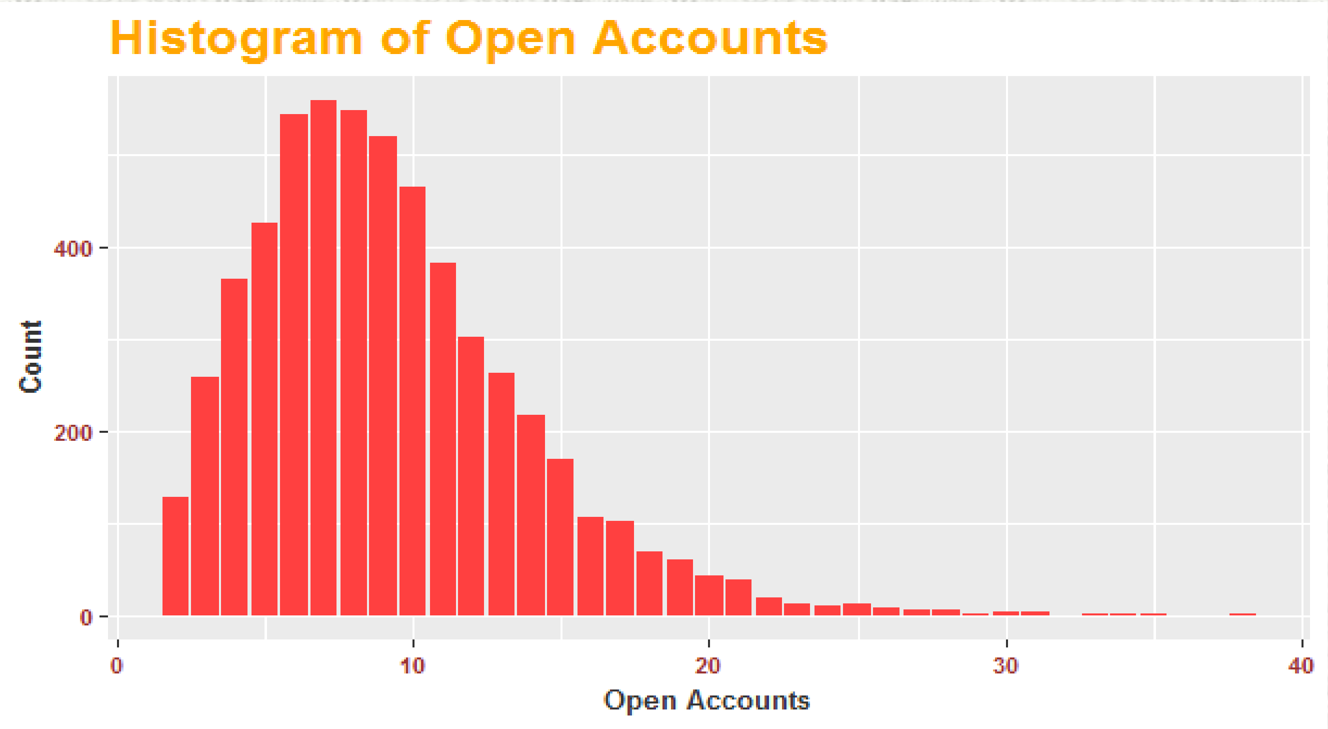
UA-Plot-11 Interest rate analysis in charged off cases



There is not much variation seen. We can see further from grade and sub-grade analysis because our understanding is that grade and subgrade determines ROI.

UA-Plot-12 Open account analysis in charged off cases

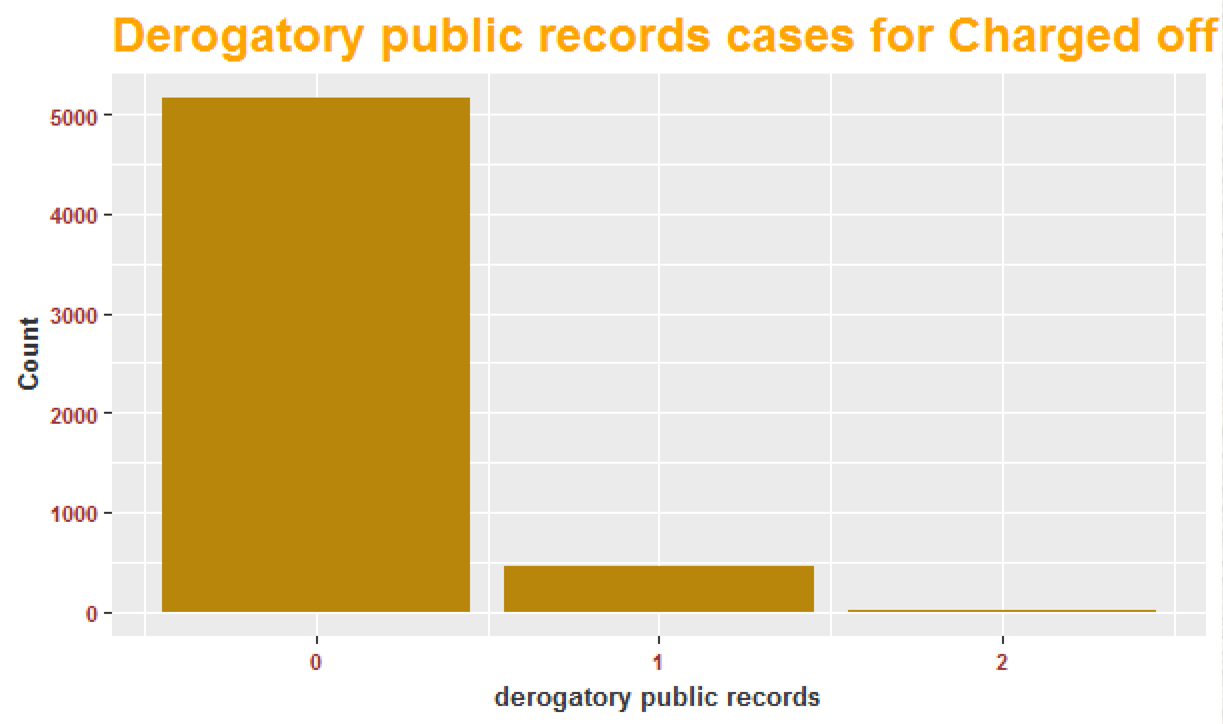


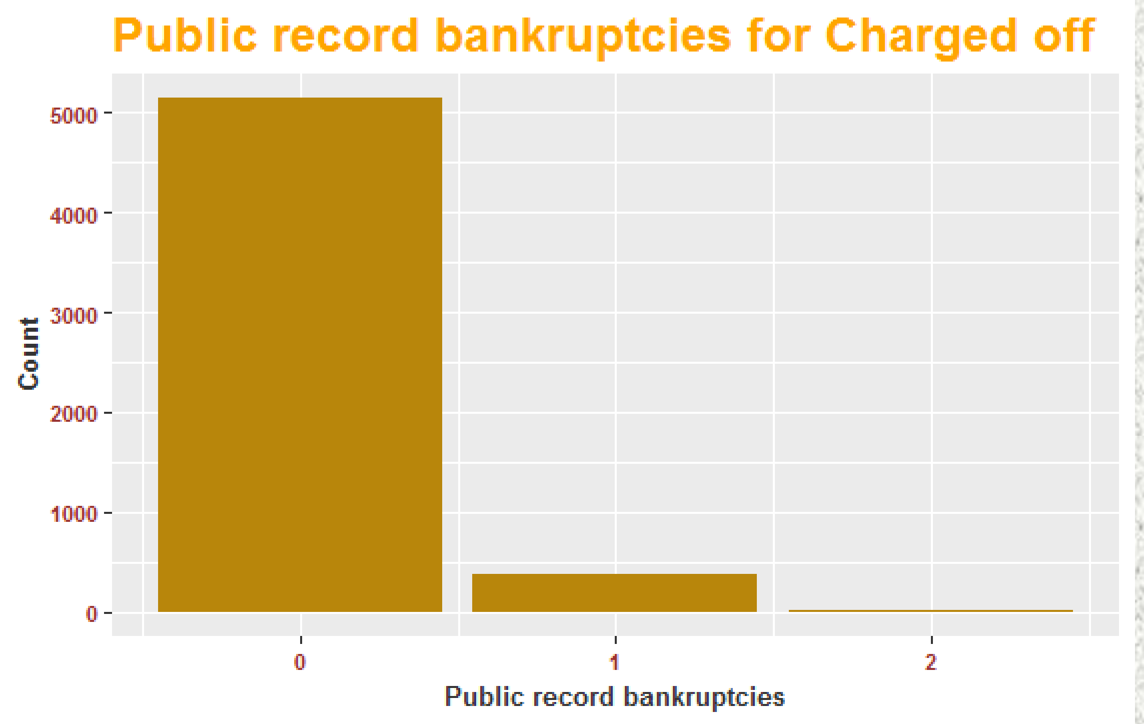


```
summary(loan_Charged_off$open_acc)
Min. 1st Qu. Median Mean 3rd Qu. Max.
2.000 6.000 8.000 9.178 12.000 38.000
```

From box plot we can see that there are many outliers which needs to be removed at time of analysis.

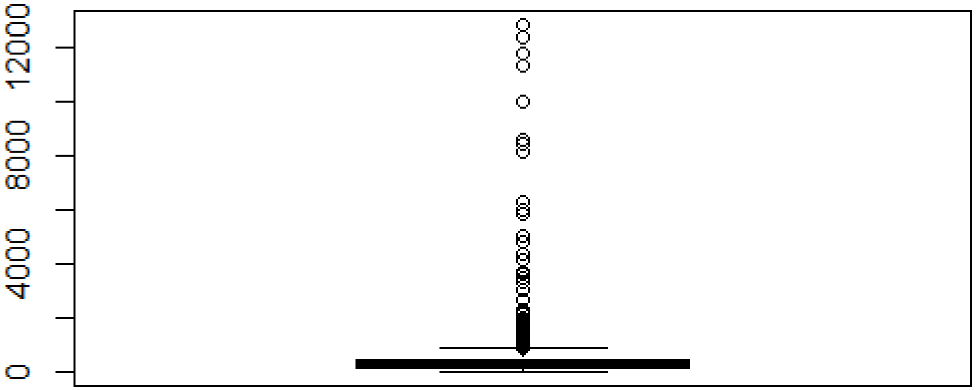
UA-Plot-13 Public Record & Public record Bankruptices in charged off cases


Derogatory public records cases for Charged off


Public record bankruptcies for Charged off

From plots we can see that there are hardly any derogatory public records or Bankruptcies public record for charged of cases. There are no NA values in pub_rec column but there are NA values in pub_rec_bankruptcies columns which needs to be cleaned.

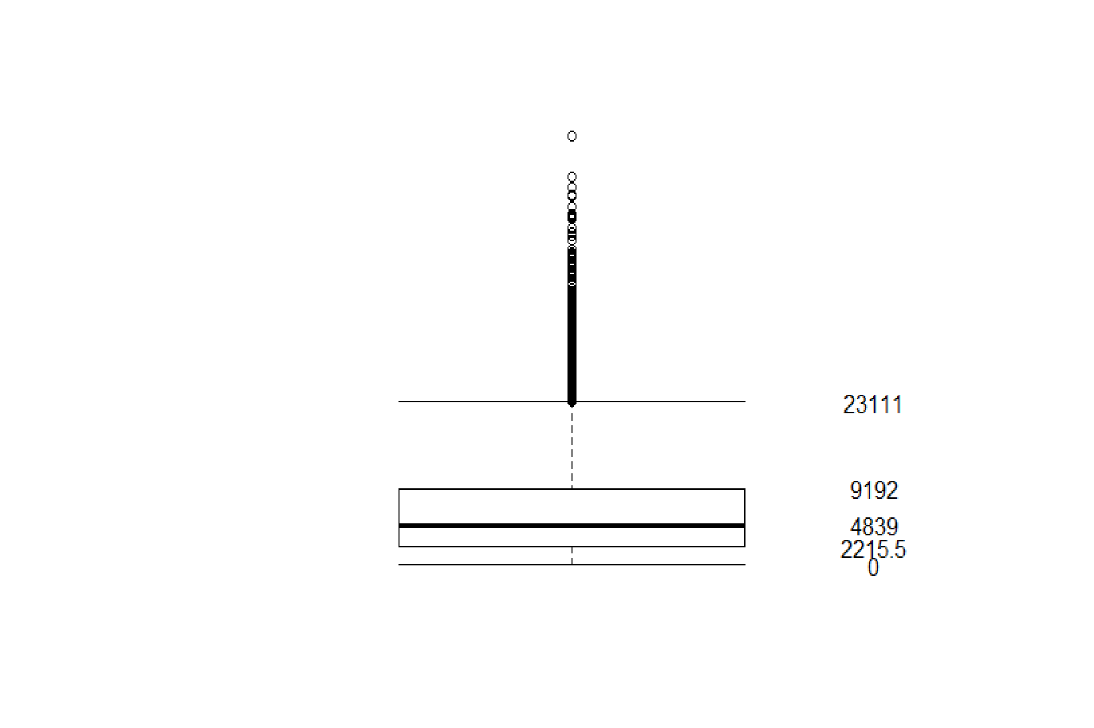
UA-Plot-14 last_pymnt_amnt distribution in charged off cases



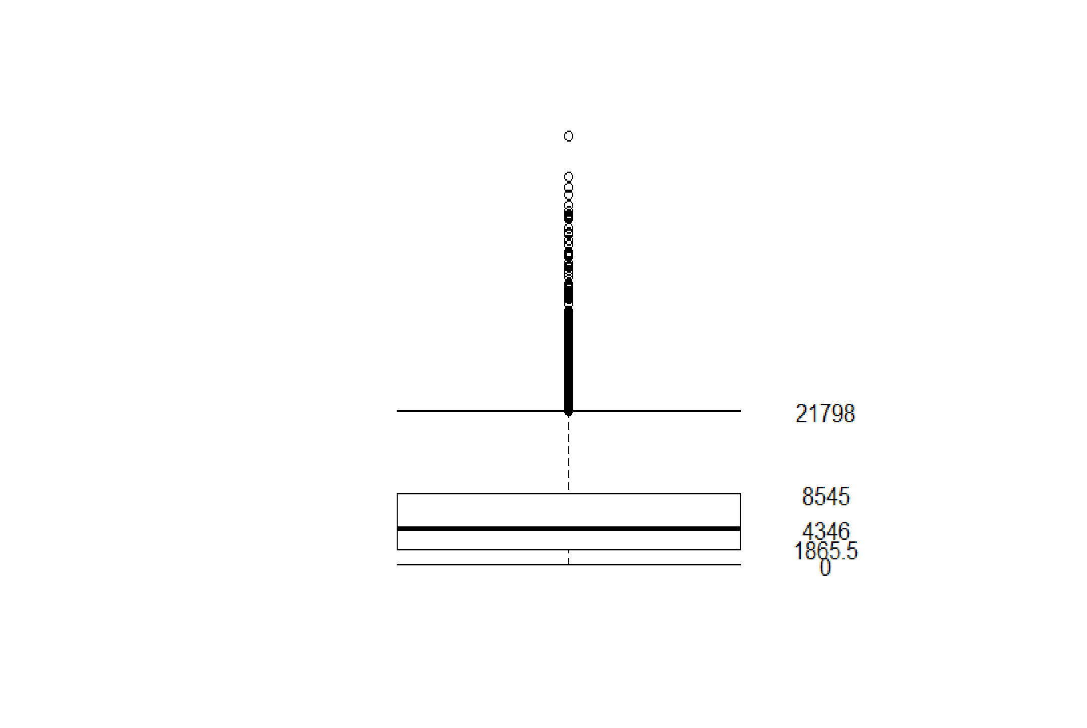
```
summary(loan_Charged_off$last_pymnt_amnt)
Min. 1st Qu. Median Mean 3rd Qu. Max.
0.0 112.9 238.2 326.0 405.7 12818.4
```

From box plot we can see that there are many outliers which needs to be removed at time of analysis.

Plot-15 Total Payment &Total Payment inv dist in charged off cases



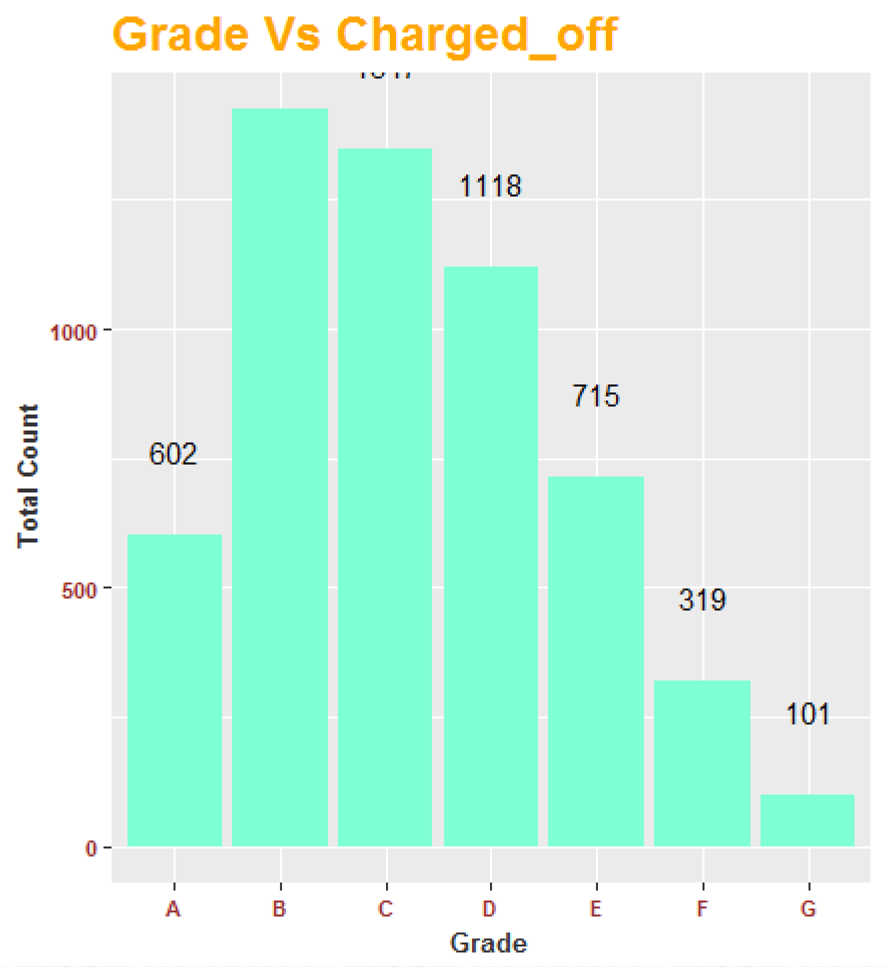
```
summary(loan_Charged_off$total_pymnt)
Min. 1st Qu. Median Mean 3rd Qu. Max.
0 2216 4839 6838 9192 51745
```

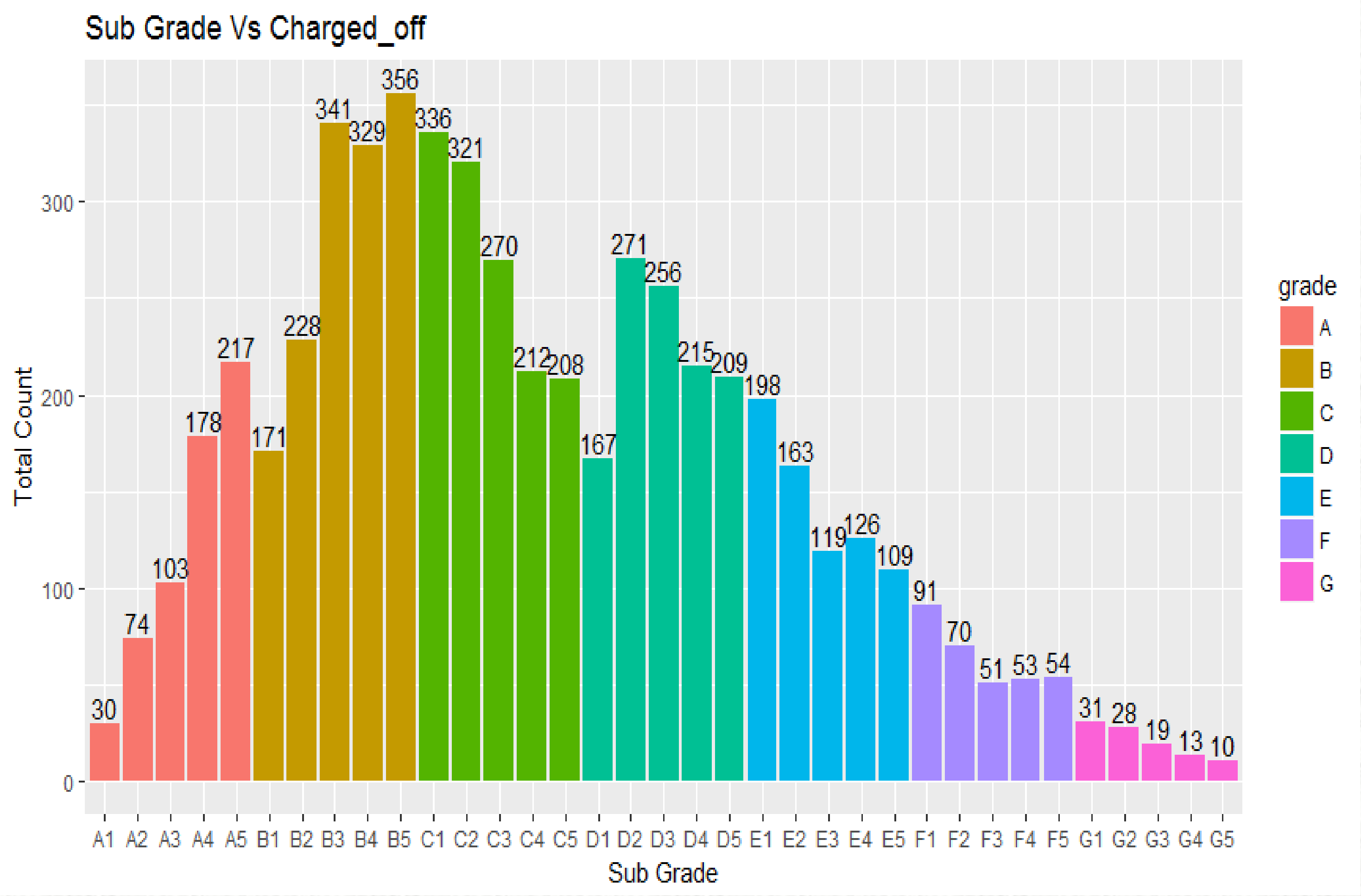


```
summary(loan_Charged_off$ total_pymnt_inv)
Min. 1st Qu. Median Mean 3rd Qu. Max.
0 4431 7907 9737 13276 55867
```

From both the box plots we can see that there are many outliers which needs to be removed at time of analysis. However both plots have similar distribution and high correlation.

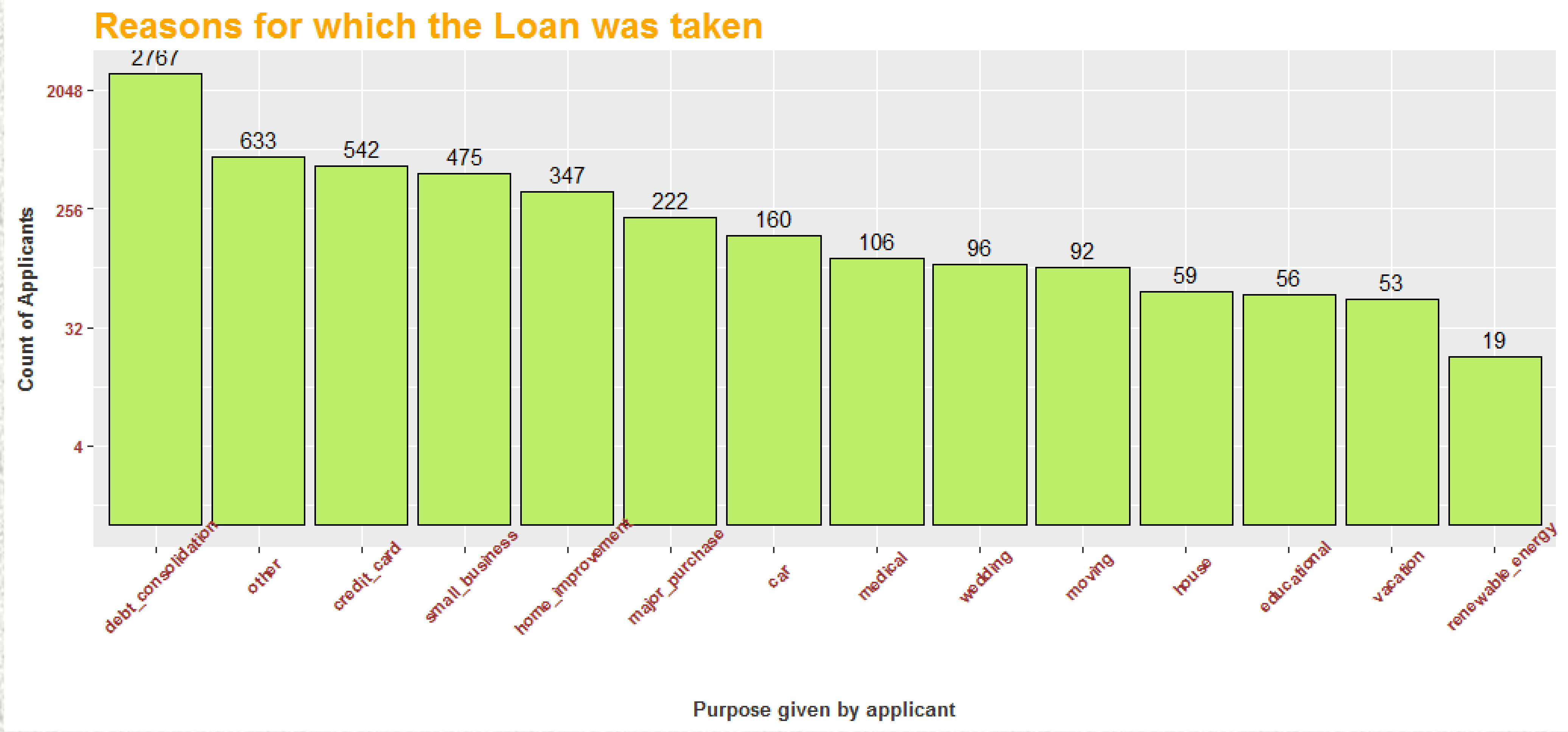
Plot-16 Grade & Sub-grade dist. in charged off cases





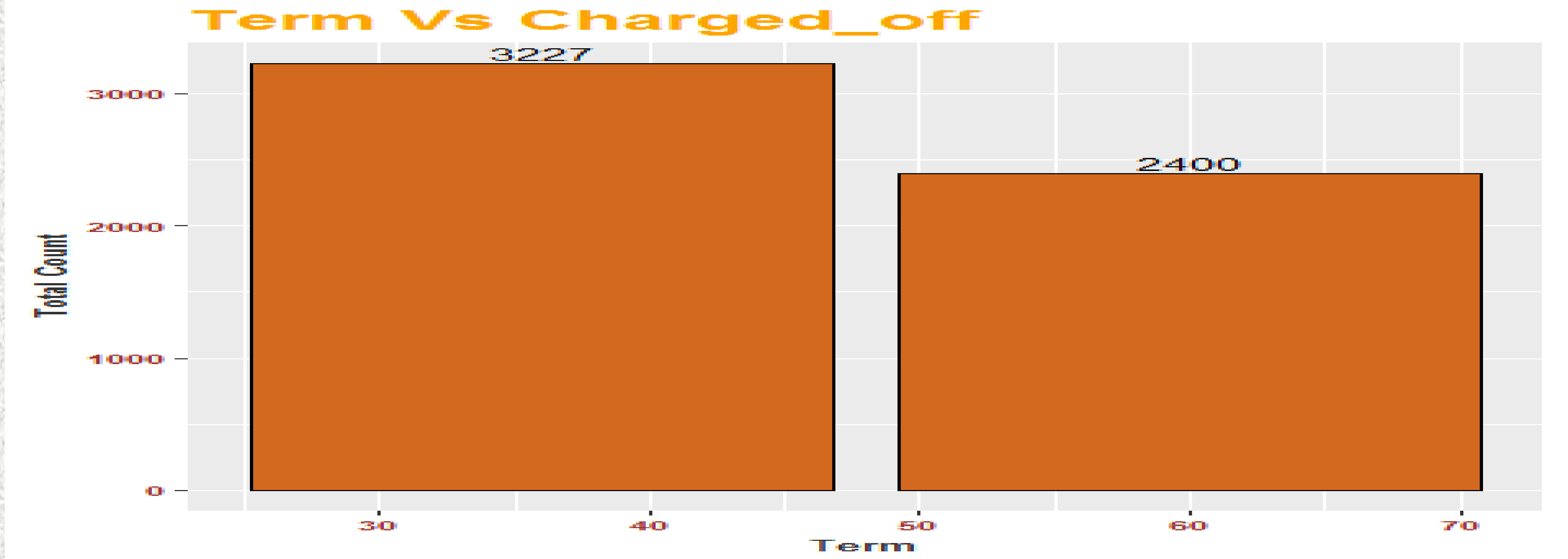
There is an evident downtrend in grade for charged off loan cases. Loan grade A to G are least risky to most risky.
It appears from plot 2 that most charged off cases are in grades B3~C3 and then from D2~E1.

UA-Plot-17 loan purpose analysis for charged off cases



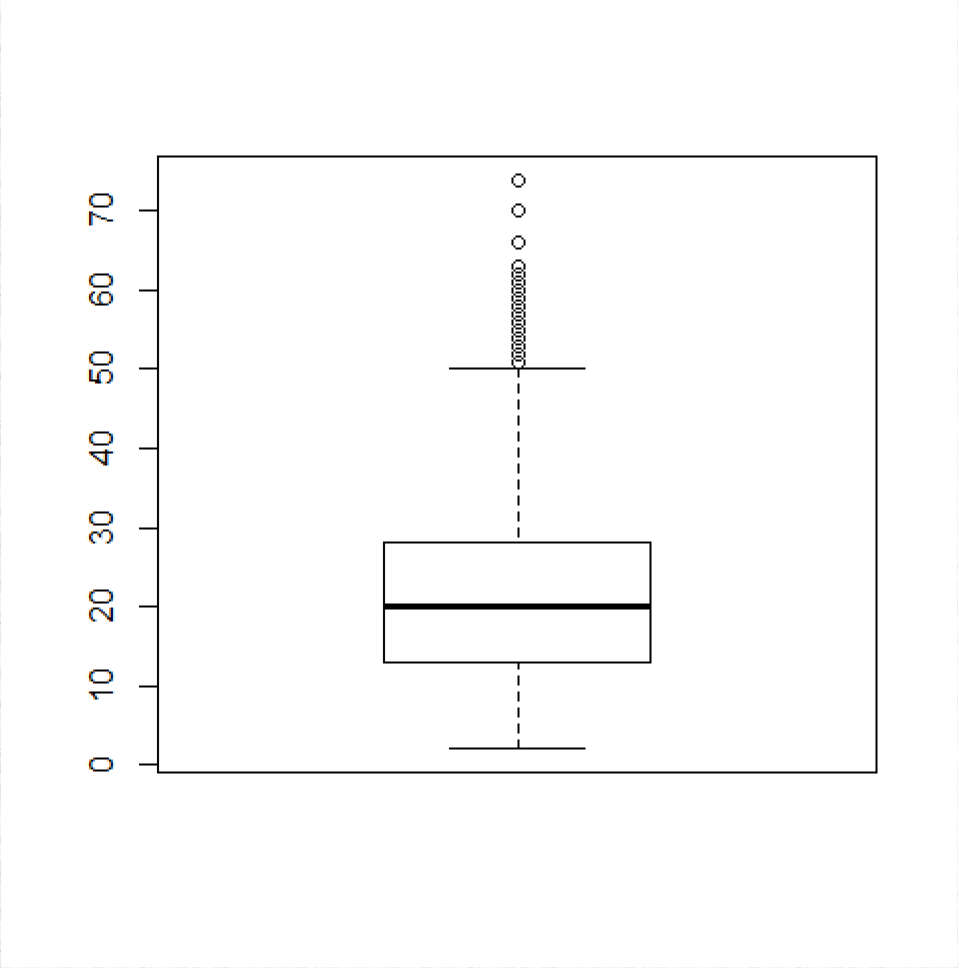
‘Debt consolidation’ have highest number of charged off cases.Second highest is other , which shows that data collection method is inadequate and we should have most categories for loan purpose.

UA-Plot-19 Term Vs Charged_off



Total charged off for 36 months tenure(3years) is greater than 60 months(5 Years).

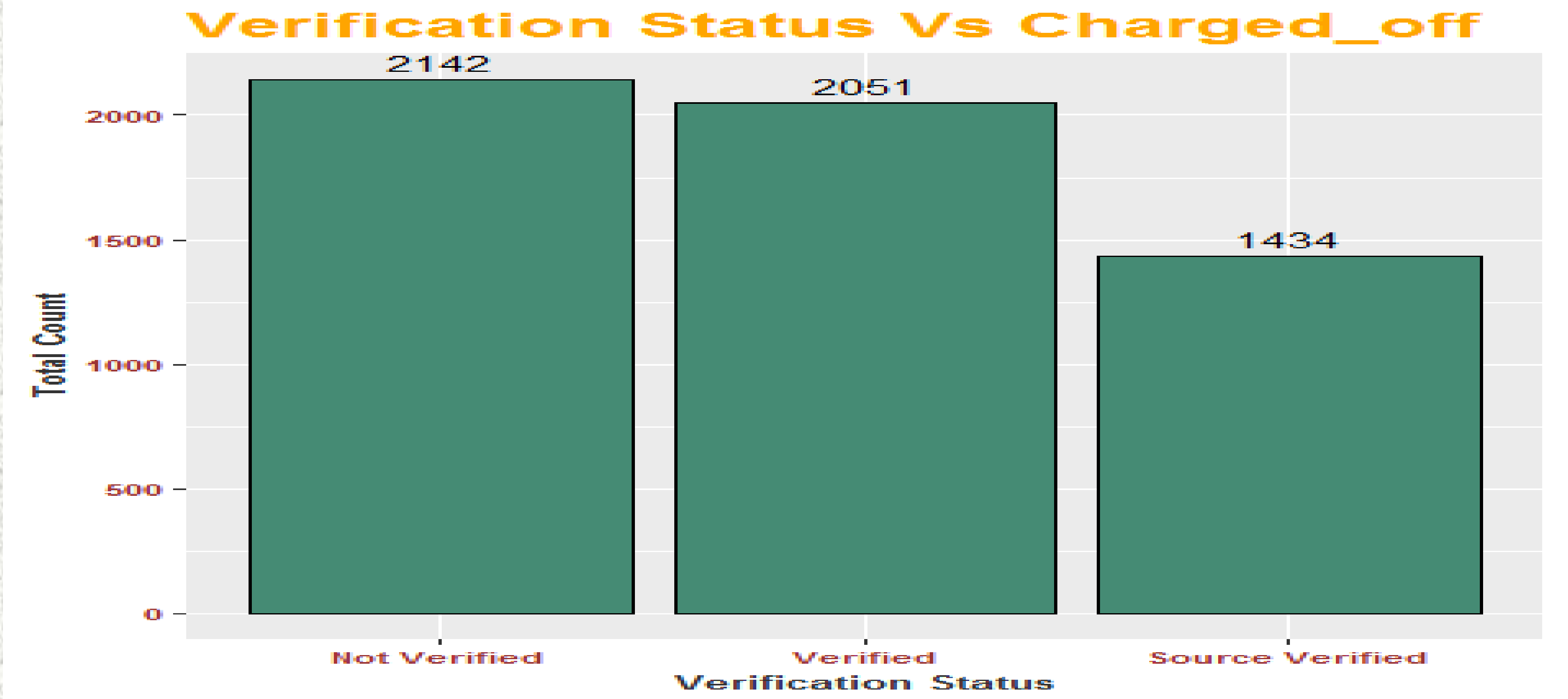
UA-Plot-20 Total open CLs for Charged_off



```
summary(loan$total_acc)
Min. 1st Qu. Median Mean 3rd Qu. Max.
2.00 12.00 18.00 20.06 26.00 87.00
```

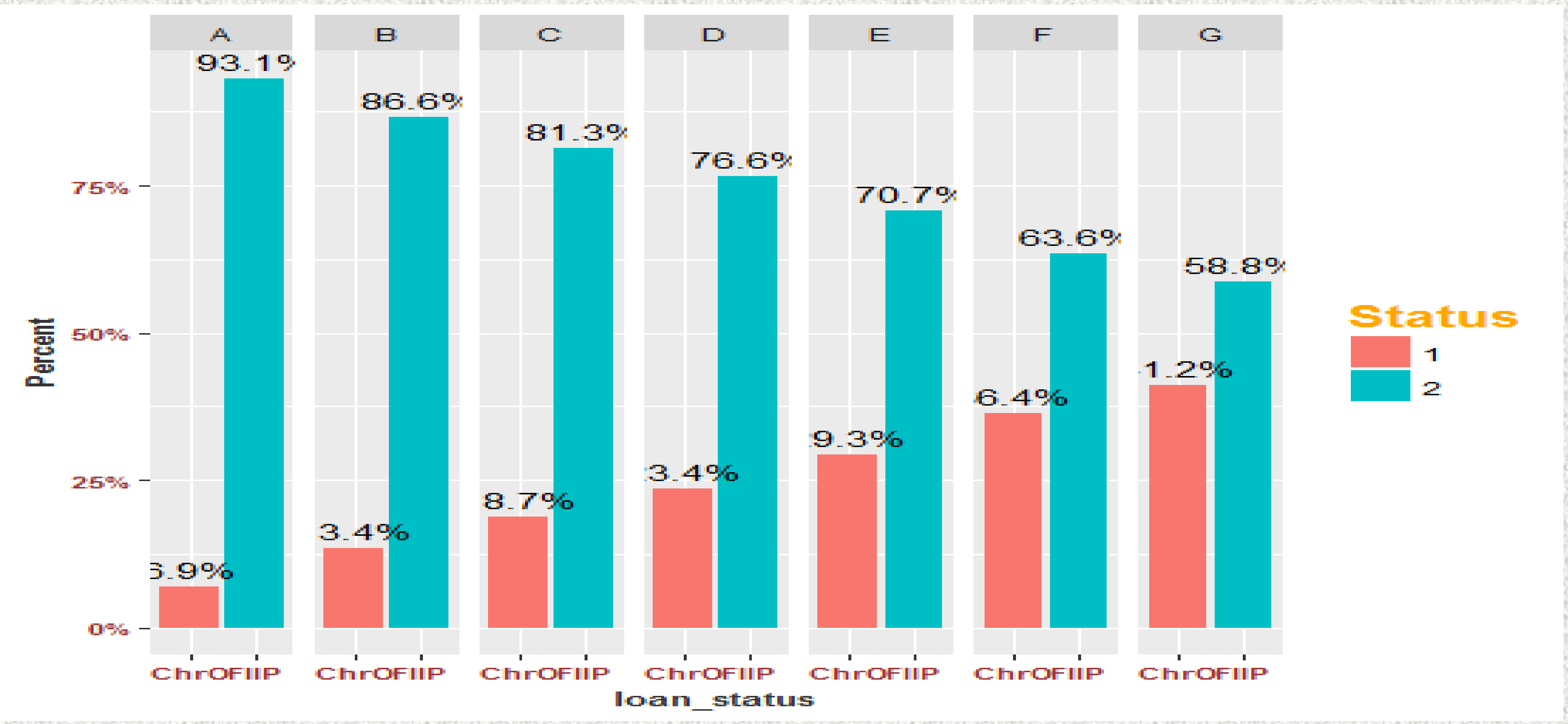
From box plot we can see that there are many outliers which needs to be removed at time of analysis.

UA-Plot-20 Verification Status Vs Charged_off



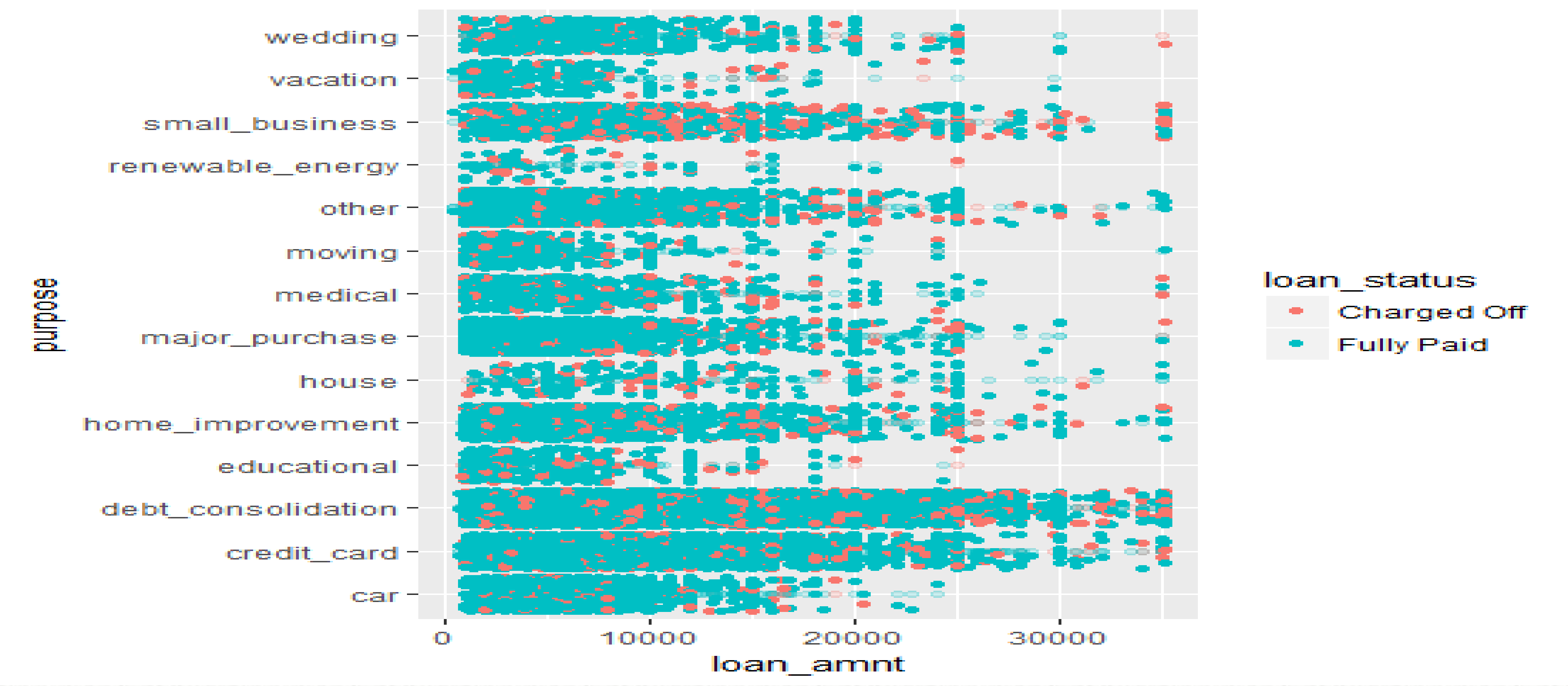
From plot it can be seen that no of charged off cases are more when income is not verified. And it is least in when income is source verified.

Bi-variate Analysis Plot-1 Grade Vs loan_status



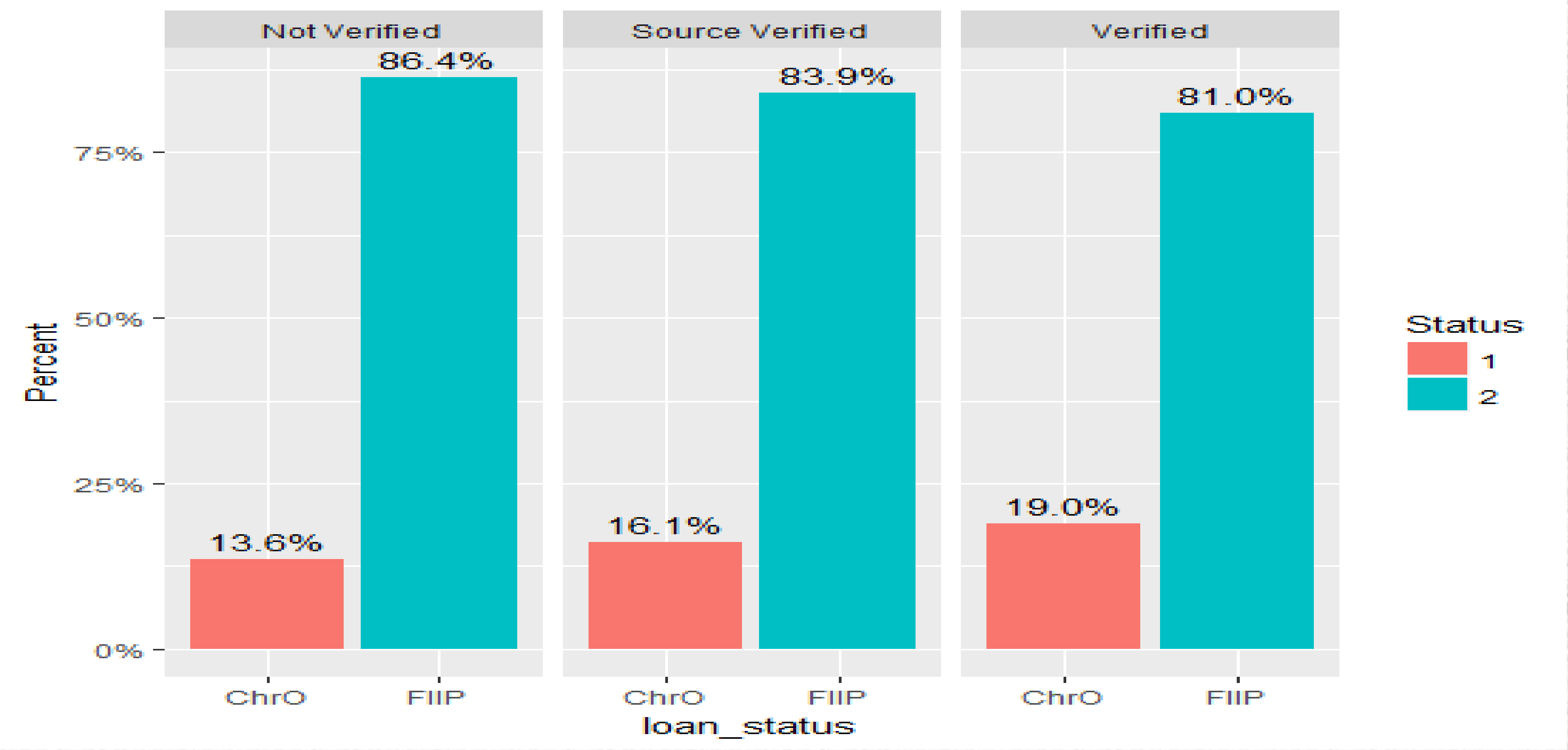
From plot it can be seen that no Grade-G has maximum number of defaulters.

Bi-variate Analysis Plot-2 Purpose Vs loan_status



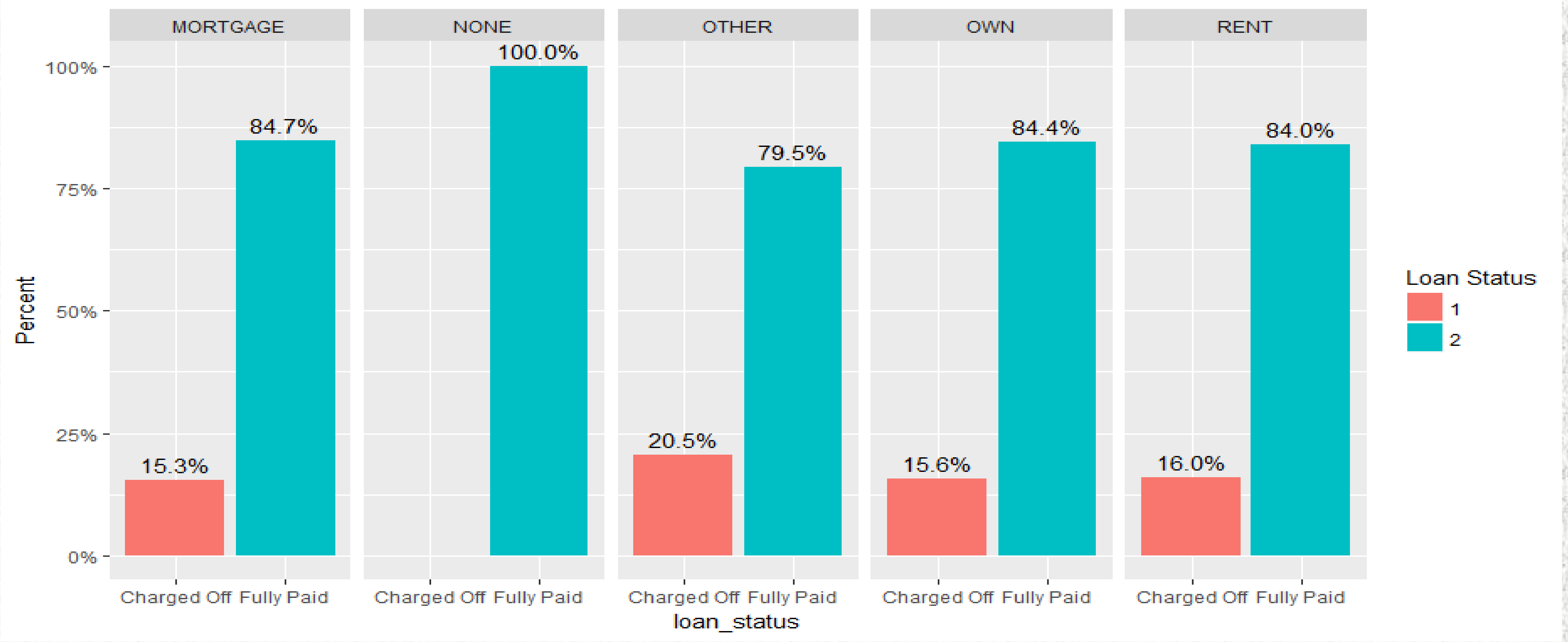
Loan borrowers with purpose as ‘debt consolidation’ have maximum chance to become defaulter.

Bi-variate Analysis Plot-3 Verification status Vs loan_status



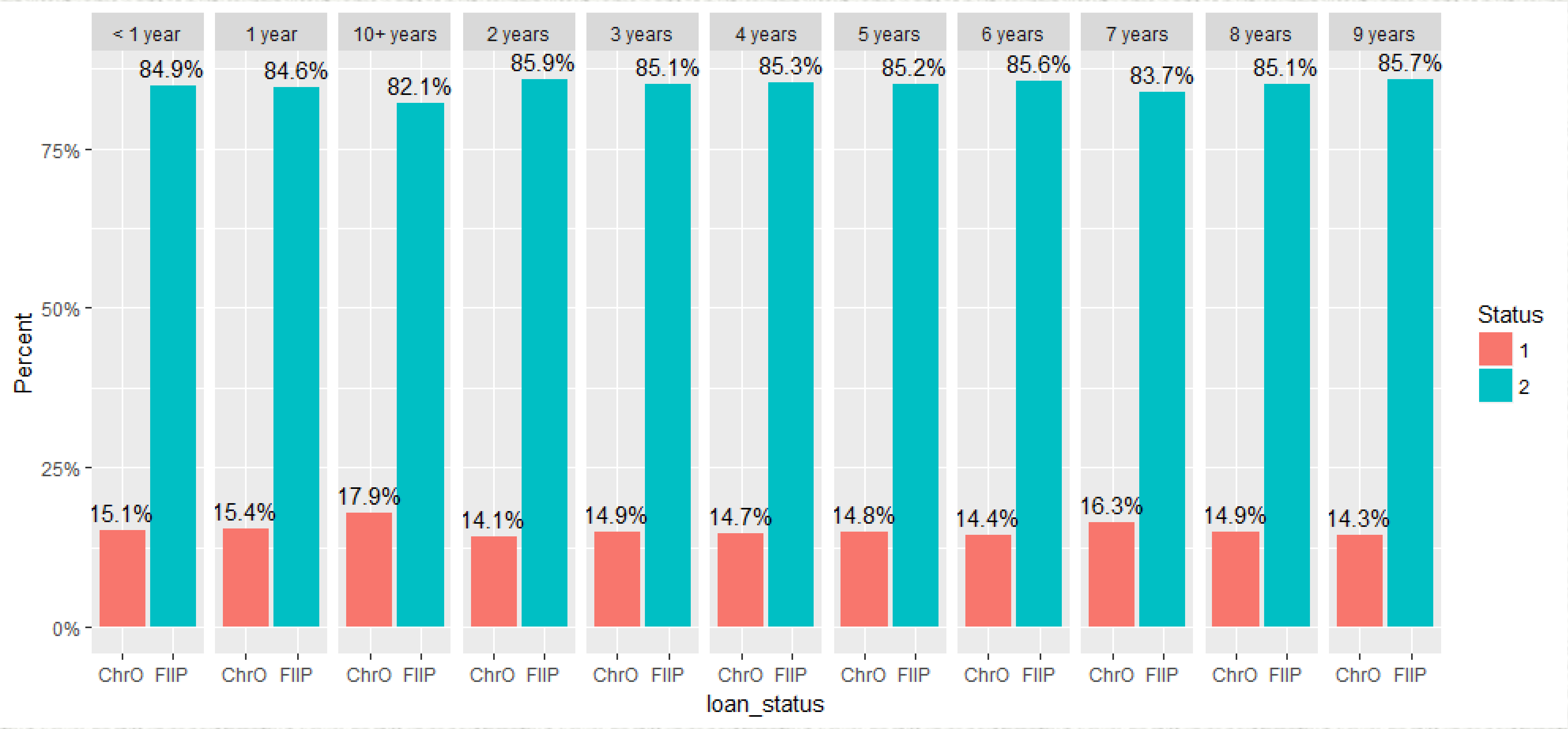
Loan borrowers defaults more in verification status as ‘Verified’.

Bi-variate Analysis Plot-4 Home Status Vs loan_status



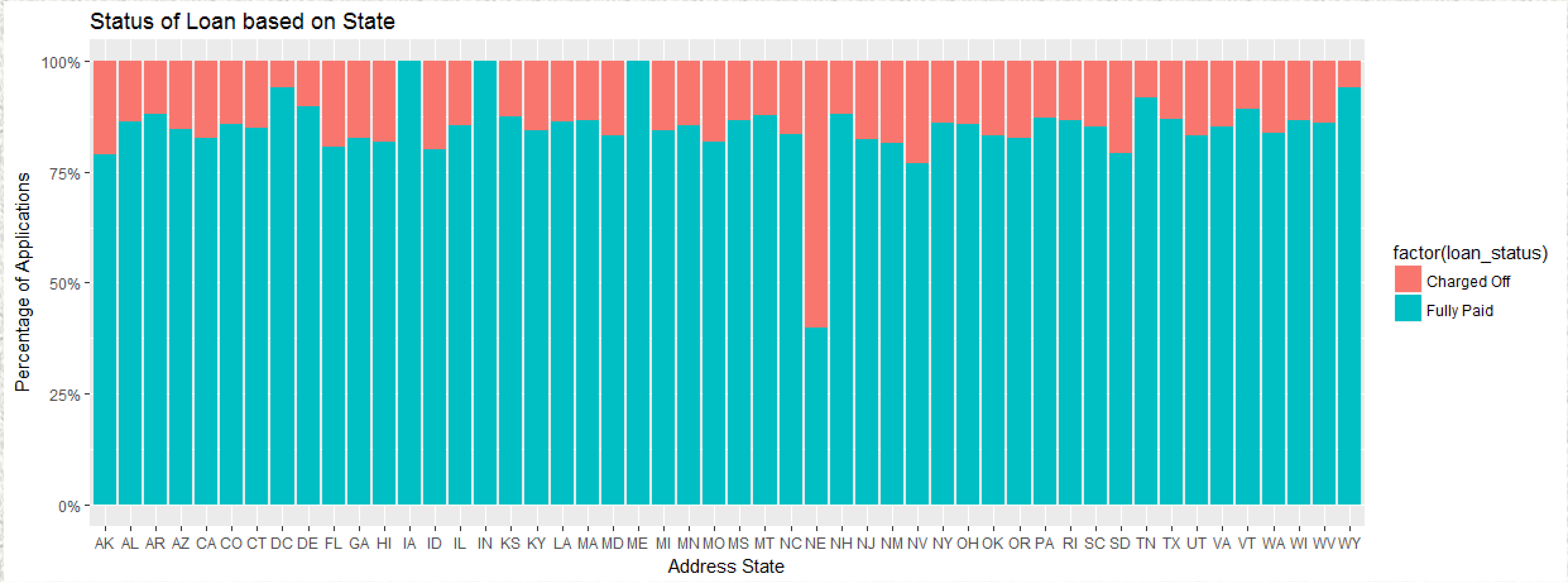
Home ownership status as ‘OTHER’ are likely to become defaulter. It looks like more data needs to be collected for proper analysis.

Bi-variate Analysis Plot-5 Employment tenure Vs loan_status



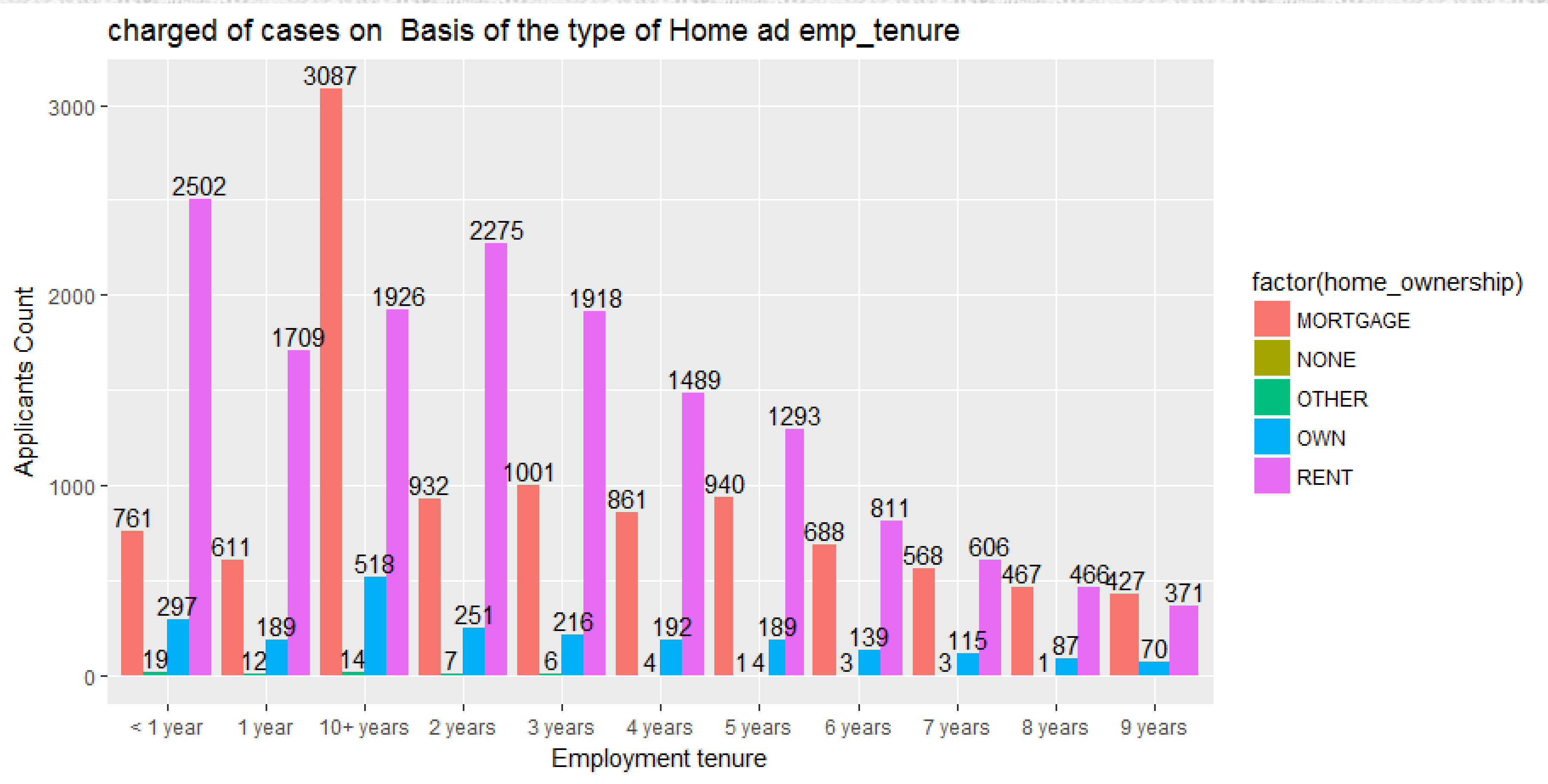
Borrowers with employment tenure >10 are most likely to default.

Bi-variate Analysis Plot-6 State Vs loan_status



Newada(NE) has most number of defaulter cases. So it is a risky state to lend loan.

Bi-variate Analysis Plot-7 Charged off on home ownership & emp_len



Maximum charged of cases are when employment tenure>10 and homeownership =“Mortgage”

Conclusion

❖ Most Defaulting Borrowers are from NEVEDA.

❖Most defaulting borrowers mention purpose of loan application as “Debt-Consolidation”

❖Most Defaulting Borrowers have “OTHERS” as ownership status. This shows data collection method is not correct and more information should be gathered at time of loan application.

❖Percentage of defaulter is highest in case when income is verified.

❖Grade G has highest number of defaulter .

Driving variables analyzed are:

❑State and zip code
❑Purpose of loan application
❑Ownership status
❑Verification status of Income
❑Grade and sub-grade
❑Employment tenure
❑Term of loan
❑Enquiry_last 6months